Sparse Video Generation Propels Real-World Beyond-the-View Vision-Language Navigation

Hai Zhang* Siqi Liang* Li Chen Yuxian Li Yukuan Xu Yichao Zhong Fu Zhang Hongyang Li
The University of Hong Kong
https://github.com/OpenDriveLab/SparseVideoNav

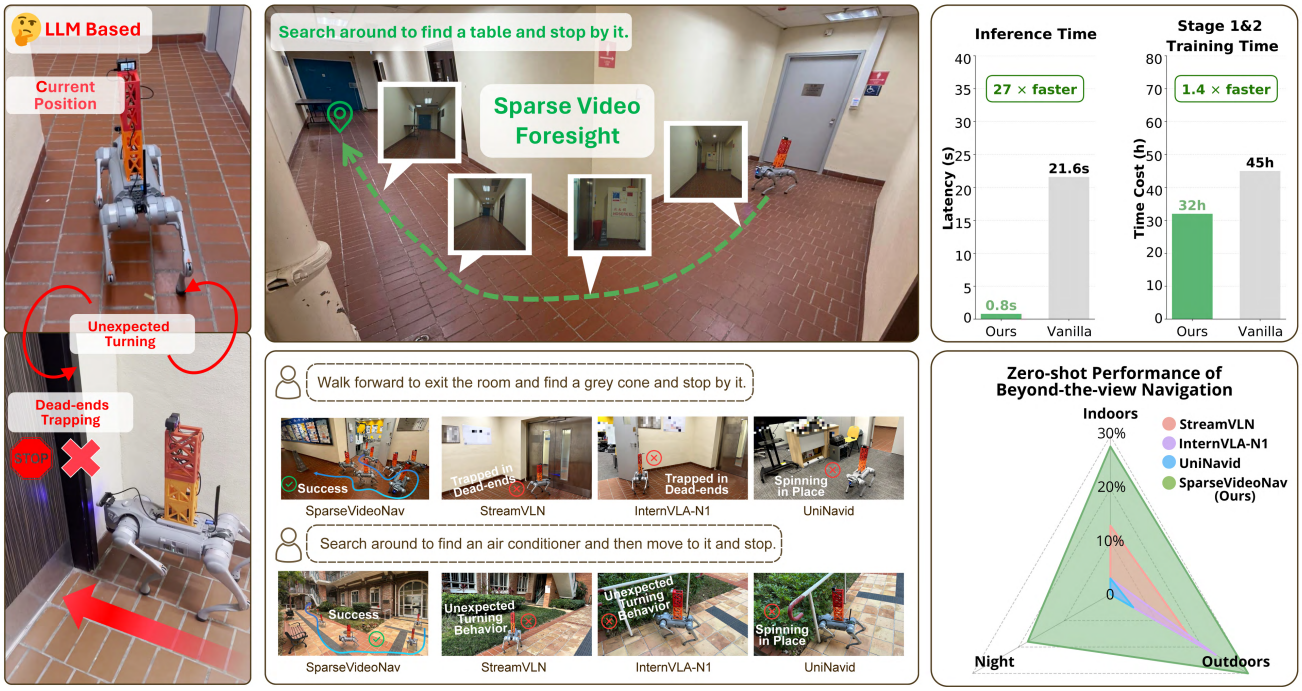

Fig. 1: In this work, we investigate the beyond-the-view navigation task in the real world, where agents must locate distant, unseen targets without step-by-step guidance. Traditional large language model-based methods suffer from short-horizon supervision, leading to short-sighted behaviors, *e.g.*, unexpected turning and dead-end trapping. We address this challenge from a new perspective, by introducing the video generation model to this field for the first time. The whole training pipeline is sparsified further for the sake of extended prediction horizon and computational efficiency.

***Abstract*—Why must vision-language navigation be bound to detailed and verbose language instructions? While such details ease decision-making, they fundamentally contradict the goal for navigation in the real-world. Ideally, agents should possess the autonomy to navigate in unknown environments guided solely by simple and high-level intents. Realizing this ambition introduces a formidable challenge: Beyond-the-View Navigation (BVN), where agents must locate distant, unseen targets without dense and step-by-step guidance. Existing large language model (LLM)-based methods, though adept at following dense instructions, often suffer from short-sighted behaviors due to their reliance on short-horizon supervision. Simply extending the supervision horizon, however, destabilizes LLM training. In this work, we identify that video generation models inherently benefit from long-horizon supervision to align with language instructions, rendering them uniquely suitable for BVN tasks. Capitalizing on this insight, we propose introducing the video generation model into this field for the first time. Yet, the prohibitive latency for generating videos spanning tens of seconds makes real-world deployment impractical. To bridge this gap, we propose SparseVideoNav, achieving sub-second trajectory inference guided by a generated sparse future spanning a 20-second horizon. This yields a remarkable 27× speed-up compared to the unoptimized counterpart. Extensive real-world zero-shot experiments demonstrate that SparseVideoNav achieves 2.5× the success rate of state-of-the-art LLM baselines on BVN tasks and marks the first realization of such capability in challenging night scenes.**

I. Introduction

Vision-language navigation (VLN) empowers agents to perform complex tasks by grounding natural language instructions into sequential actions based on visual observations [19]. Recently, the advent of large-language models (LLMs) has catalyzed significant breakthroughs in this field [49, 50, 41]. However, a fundamental tension remains: while real-world interactions typically demand navigation based on simple and high-level intents, current agents often require dense and step-by-step instructions [8, 42], a paradigm typically termed Instruction-Following Navigation (IFN). This clear discrepancy highlights a formidable but less-explored fron-

tier: *Beyond-the-View Navigation* (BVN), where agents must autonomously locate unseen targets in the absence of such granular and intermediate guidance.

The primary bottleneck in BVN for existing LLM-based methods is their inherent short sight. These models are typically supervised by short-horizon actions (*e.g.*, 4 to 8 steps) during training [42, 49, 35]. Consequently, they struggle to infer correct navigational intents, typically causing two failure modes during deployment. First, the inability to observe the target over a long distance induces severe uncertainty, resulting in unexpected turning behaviors or spinning in place. Second, upon entering a dead end, the agent mistakenly assumes the path end, leading to dead-end trapping. While extending the supervision horizon seems like a logical fix, it often destabilizes the training process of LLM [40], rendering it a non-viable solution.

In this work, we pivot toward video generation models (VGMs) based on a key observation: unlike LLMs, VGMs are inherently pre-trained to capture long-horizon future aligned with the language instruction [38, 43]. Recognizing this strength, we identify VGMs as a well-suited interface to provide the long-horizon foresight that BVN demands. Nevertheless, whether we should strictly adhere to the standard VGM paradigm of predicting ***continuous*** videos [14, 35] remains a question worthy of scrutiny. We contend that the dense and high-frequency temporal information required by continuity is redundant for guiding navigation. This insight drives us to explore a ***sparse*** video generation paradigm. By doing so, we aim to achieve the dual advantage of extending the prediction horizon while simultaneously reducing both training and inference overhead. To this end, we reformulate the training objective by employing sparse video frames corresponding to strategically selected timesteps as the direct supervision signal for VGM.

Translating this sparse foresight into a comprehensive navigation system, yet, presents two challenges. Primarily, the computational overhead of injecting the whole history, coupled with the extensive denoising steps required to generate future predictions, still leads to significant inference latency that makes real-world deployment impractical. In addition, unlike LLM-based methods that can bridge the sim-to-real gap by co-training with heterogeneous real-world data [49, 42], VGM lacks a straightforward mechanism to leverage such data sources.

To bridge this gap, we introduce SparseVideoNav, a **Sparse Video** generation-based system for vision-language **Nav**igation. For the efficiency bottleneck, we instantiate SparseVideoNav via a structured training pipeline. For the data challenge, we curate a large-scale navigation dataset spanning 140 hours across diverse real-world environments. Through zero-shot evaluations in six real-world scenes, we demonstrate that SparseVideoNav achieves state-of-the-art (SOTA) performance on BVN tasks, marking the first realization of such capability in challenging night scenes.

Contributions:

1) We investigate beyond-the-view navigation tasks in the real world from a new perspective by applying video generation model as the primary backbone for action generation. Crucially, we empirically identify that supervising with longer horizon is instrumental in improving performance.

2) We advance beyond the conventional constraint of continuous video generation in existing methods by pioneering a paradigm shift towards sparse video generation. This reformulation liberates the capability to reason over substantially longer prediction horizon while achieving $1.4\times$ training speed-up and $1.7\times$ inference speed-up.

3) SparseVideoNav achieves sub-second trajectory inference guided by a generated sparse future spanning a 20-second horizon. This yields a remarkable $27\times$ speed-up compared to the unoptimized counterpart. Extensive real-world zero-shot experiments demonstrate SparseVideoNav adeptly handles beyond-the-view tasks even in complex terrains like dead ends, narrow accessible ramp, and hillside with high inclination angles, achieving $2.5\times$ the success rate of SOTA LLM baselines and marking the first realization of such capability in challenging night scenes with a 17.5% success rate.

II. Related Work

A. Foundation Models for Vision-Language Navigation

The integration of foundation models into VLN predominantly transitions from modular, training-free approaches to end-to-end fine-tuning approaches. Training-free approaches leverage the zero-shot reasoning capabilities of off-the-shelf LLMs to perform specific sub-tasks, such as task reasoning [24], stage scheduling [6], frontier detection [11], and object validation [48]. While these modular approaches maintain interpretability, they inherently suffer from cascading error propagation [48] or limited generalization capability [47, 7].

To better optimize spatio-temporal information throughout the training pipeline, fine-tuning approaches have gained increasing traction, which aim to map multimodal inputs directly to actions in an end-to-end manner [49, 51]. Recent SOTA methods have demonstrated strong generalization through large-scale training. Notably, Zhang et al. [49] proposes Uni-Navid to establish a competitive baseline by unifying diverse navigation tasks with a single versatile policy. StreamVLN [42] introduces a streaming framework that manages continuous video inputs via hybrid slow-fast context modeling to accelerate the inference. Despite their promise, these methods predominantly require dense and step-by-step instructions, a dependency that contradicts with practical real-world demands. When facing beyond-the-view tasks guided by simple and high-level intents, they are prone to falter due to the vulnerability induced by short-horizon supervision. Consequently, there is a pressing need for a novel paradigm designed to bridge this gap.

B. Video Generation Models for Embodied Agents

Distinguished from text-centric pretraining, VGM implicitly encodes dynamic variation between frames, leading to a smaller domain gap towards downstream actions. This unique property has driven an explosion of interest across a wide spectrum of embodied tasks from robotic manipulation [3, 14, 46]

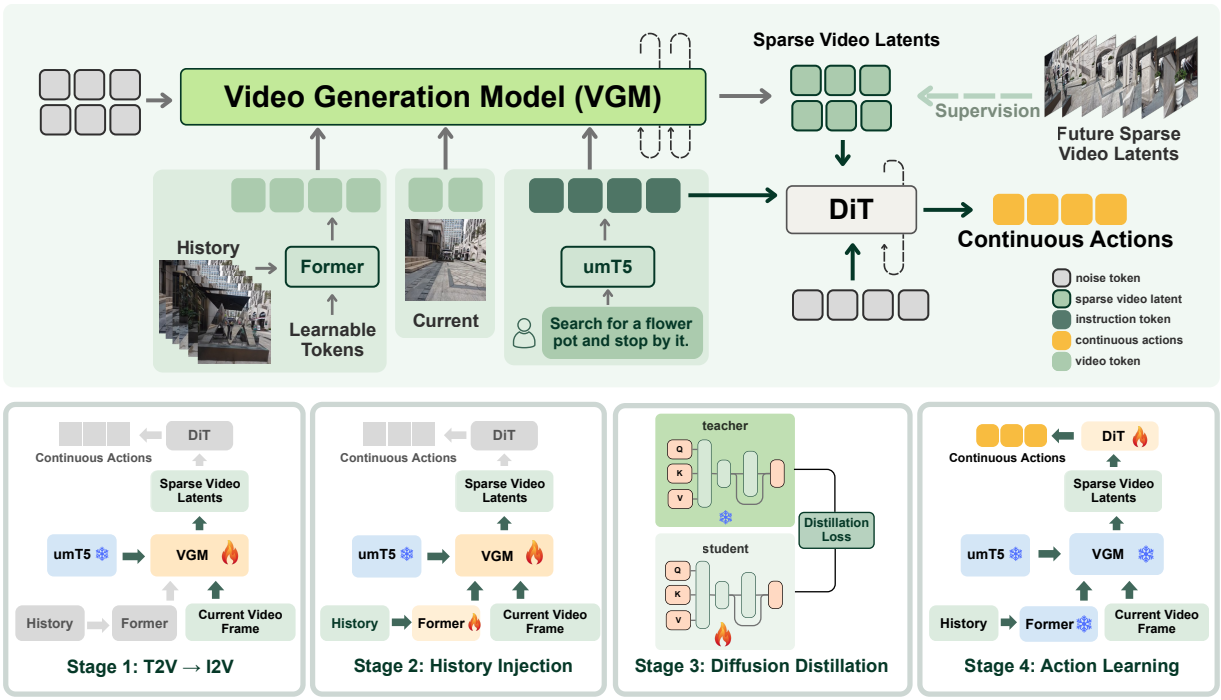

Fig. 2: **Architecture and four-stage training pipeline of SparseVideoNav**. **(Top)** denotes our whole training architecture. Current observation, historical observations, and the language instruction are fed into the video generation model (VGM) backbone to generate future sparse video latents. DiT-based action head then predicts continuous actions conditioned on generated sparse future and the language instruction. **(Bottom)** denotes our four-stage training pipeline, with **Stage 1 (Sec. III-C)** adapting T2V to I2V, **Stage 2 (Sec. III-D)** injecting history into I2V backbone; **Stage 3 (Sec. III-E)** distilling the backbone to reduce denoising steps; **Stage 4 (Sec. III-F)** learning actions based on generated sparse future. Components not utilized in a specific stage are indicated by gray blocks.

to autonomous driving [10, 45, 52]. Within the specialized scope of navigation, InternVLA-N1 [35] demonstrates that co-training with video generation objectives can improve downstream policy performance. However, its primary backbone remains LLM, making it susceptible to the limitations of short-horizon supervision. To the best of our knowledge, SparseVideoNav is the first to apply VGM as the primary backbone for action generation in VLN. Unlike Navigation World Model [1], which is restricted to predicting the immediate next frame based on low-level action inputs, our framework is purely language-conditioned, directly processing high-level intents to achieve long-horizon video foresight.

Noticeably, these endeavors adopt the default prerequisite to generate continuous videos. While some concurrent works in manipulation explore sparsification to balance action and video tokens [2, 21], their future predictions are typically limited to a mere few seconds. In contrast, we are the first to harness the sparsification for significantly extending the prediction span (*e.g.*, 20 seconds). This enables efficient long-horizon supervision, ultimately unlocking substantial performance improvements in BVN tasks.

III. Methodology

The design of SparseVideoNav is guided by two fundamental pillars: a training paradigm-level innovation through sparsification (Sec. III-A), and a system-level innovation via synergistic orchestration. To achieve this, we develop a data curation pipeline (Sec. III-B) alongside a structured four-stage training pipeline. The linchpin of the training pipeline is the *sparse future video*, which serves as the supervision across stages 1, 2, and 3, and the conditional input of stage 4.

The whole architecture and training pipeline are shown in Fig. 2, where `Former` block denotes Q-Former [20] and Video-Former [14] used to compress history.

A. Sparsification

To extend the prediction horizon, we bring in sparse video supervision into the training of VGM to enable fixed-interval video generation. As shown in Fig. 3, setting the interval to 3 strikes the best balance between the prediction horizon and the visual fidelity. To further ensure action prediction accuracy, we maintain continuous generation for the first two observation chunks, covering 8 timesteps. This design leads to the sparse generation timesteps at $[T+1, T+2, T+5, T+8, T+11, T+14, T+17, T+20]$, spanning 20 seconds at 4 FPS, where T denotes the current timestep.

B. Data Curation Pipeline

Different from methods built upon video-based LLMs [49, 50], which can mitigate the sim-to-real gap by co-training on pure simulation navigation data [28] mixed with real-world VQA datasets, VGM cannot effectively leverage this

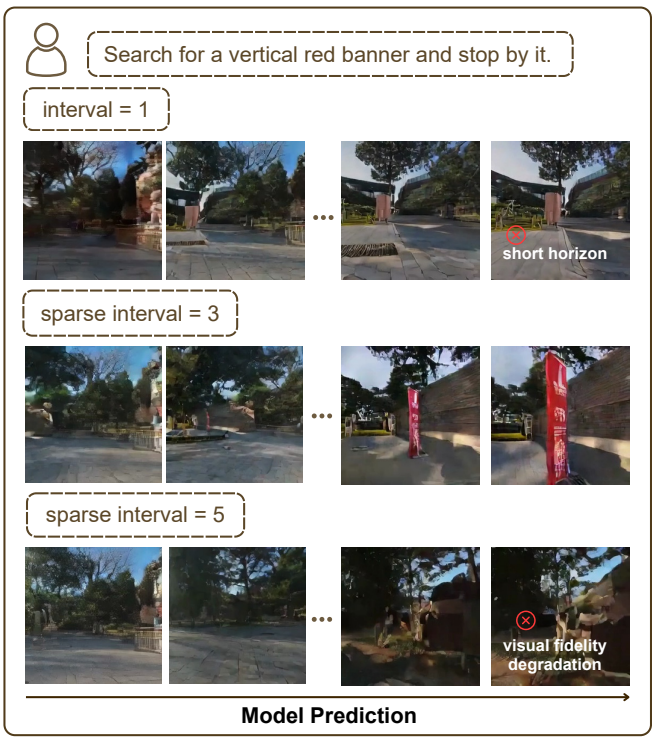

Fig. 3: **Qualitative comparison of different sparse intervals.** With the sparse interval of 3, the model successfully imagines a path towards beyond-the-view target, while maintaining visual fidelity.

strategy. Relying exclusively on simulation data typically leads to mode collapse [34, 36]. Furthermore, existing real-world navigation datasets are often characterized by severe fisheye distortion [29, 13] and limited scale [12], making them unsuitable for fine-tuning VGM.

To address these challenges, we develop a data curation pipeline involving human operators equipped with a handheld camera to capture diverse videos. To minimize human jitter that blocks VGM to learn consistent dynamics, we explicitly utilize DJI Osmo Action 4 with RockSteady+ stabilization.

We collect 140 hours of real-world navigation videos. These videos are processed into approximately 13,000 trajectories via uniform temporal sampling, with an average length of 140 frames at 4 FPS. Subsequently, we estimate camera poses using Depth Anything 3 (DA3) [22] to extract continuous action labels. Detailed visualizations concerning action extraction are deferred in Sec. C. Finally, language instructions are manually annotated by human experts. This pipeline establishes the largest real-world VLN dataset to date. We will release this dataset to benefit the community.

C. Stage 1: T2V → I2V

To achieve a balance between computational overhead and video generation fidelity, we adopt Wan2.1-1.3B T2V [38] model as our backbone. Wan utilizes a 3D causal VAE [17] structure to compress the spatio-temporal dimensions. Specifically, given an input video $V \in \mathbb{R}^{(1+T)\times H\times W\times 3}$, Wan-VAE encodes it into latent chunks $C \in \mathbb{R}^{1+T/4,H/8,W/8,16}$, resulting in $(1+T/4)$ chunks where the shape of each chunk is $[H/8, W/8, 16]$. By default, all subsequent training processes are performed at the chunk level.

Note that the T2V model is designed to generate the future mainly based on language instructions rather than visual inputs. Hence, the first stage involves adapting the backbone from T2V to image-to-video (I2V) to ensure the consistency of the generated future with the initial observation.

We retain the original flow matching [23] objective of Wan for fine-tuning. During training, given a chunk latent at an arbitrary timestep c_T, the following sparse chunk latents $x_1 = [c_{T+1}, c_{T+2}, c_{T+5}, c_{T+8}, ..., c_{T+20}]$, a random noise $x_0 \sim \mathcal{N}(0, I)$, and a timestep $t \in [0, 1]$ sampled from a logit-normal distribution, an intermediate latent x_t is obtained as the training input. x_t is defined as a linear interpolation between x_0 and x_1:

$$x_t = tx_1 + (1-t)x_0. \tag{1}$$

The ground-truth velocity v_t is defined as:

$$v_t = \frac{dx_t}{dt} = x_1 - x_0. \tag{2}$$

The loss function is formulated as mean square error between the model output and the velocity v_t:

$$\mathcal{L}_{\text{stage1}} = \mathbb{E}_{x_0,x_1,l,c_T,t}||u(x_t, l, c_T, t; \theta) - v_t||^2, \tag{3}$$

where l is the language embedding of umT5 [9], θ is the model weights, and $u(x_t, l, c_T, t; \theta)$ denotes the predicted velocity.

D. Stage 2: History Injection

A critical distinction between navigation foundation models and previous vision-language-action models lies in the necessity of incorporating the entire history of observations [16, 5, 15]. Unlike LLMs that can directly absorb a long sequence of image tokens, VGMs lack this capability to process this input [38, 43]. We draw inspiration from CDiT [1] architecture to inject the history information with efficient training and inference computational overhead. Specifically, we introduce an additional cross-attention block within each transformer block of Wan backbone to explicitly inject history information. To preserve the generative priors of the fine-tuned I2V model, we initialize the final linear layer of these newly added cross-attention blocks with zeros.

Nevertheless, the whole history input presents another challenge due to its excessive length and high dimensionality. To efficiently extract spatio-temporal features, we employ a two-step strategy that utilizes a Q-Former [20] to process features along the temporal dimension and subsequently applies a Video-Former [14] for features along the spatial dimension.

We denote the history embedding processed by Q-Former and Video-Former at an arbitrary timestep as h_T, the training objective is formulated as:

$$\mathcal{L}_{\text{stage2}} = \mathbb{E}_{x_0,x_1,l,c_T,h_T,t}||u(x_t, l, c_T, h_T, t; \theta) - v_t||^2. \tag{4}$$

TABLE I: **Quantitative results of zero-shot performance on diverse real-world scenes.** SparseVideoNav achieves unanimously SOTA performance across all the scenes on both instruction-following navigation (IFN) and beyond-the-view navigation (BVN) tasks. The highest success rate across SparseVideoNav and all baselines is highlighted in **bold**. (Numbers represent **success rate** in %)

	Indoors				Outdoors				Night				**Average**	
	Room		Lab Bldg		Yard		Park		Square		Mountain		Total	
Method	IFN	BVN	IFN	BVN	IFN	BVN	IFN	BVN	IFN	BVN	IFN	BVN	IFN	BVN
Baselines														
Uni-NaVid [49]	20	5	10	0	10	0	20	10	0	0	0	0	10.0	2.5
StreamVLN [42]	45	10	40	15	30	5	50	30	20	0	**25**	0	35.0	10.0
InternVLA-N1 [35]	25	5	15	0	35	5	30	**40**	0	0	0	0	17.5	8.3
SparseVideoNav	**55**	**25**	**55**	**30**	**50**	**20**	**65**	**40**	**50**	**20**	**25**	**15**	**50.0**	**25.0**
Ablation Study														
(a) *w distilled 4 steps cont. 2*	25	0	20	0	10	0	35	15	5	0	0	0	15.8	2.5
(b) *w distilled 4 steps cont. 10*	50	10	40	5	35	5	45	40	30	5	20	5	36.7	11.7
(c) *w/o distilled 50 steps cont. 20*	70	35	65	45	55	30	75	60	75	20	35	25	62.5	35.8
(d) *SparseVideoNav w/o Former*	40	15	60	25	40	20	60	45	45	20	25	10	45.0	22.5

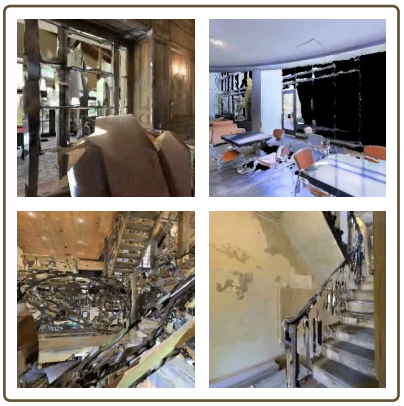

Fig. 4: **Scenes of Habitat.**

TABLE II: **Quantitative results of all methods on R2R-CE subset.**

Method	R2R Val-Unseen			
	NE↓	OS↑	SR↑	SPL↑
Uni-NaVid [49]	5.29	52.36	42.31	38.19
InternVLA-N1 [35]	4.72	60.37	51.95	45.59
StreamVLN [42]	4.97	60.64	50.01	41.50
SparseVideoNav	**3.80**	**62.62**	**53.34**	**50.21**

E. Stage 3: Diffusion Distillation

Unlike manipulation tasks, which involve limited visual changes [44, 16, 30] and allow high-fidelity reconstruction with few denoising steps [14, 4], navigation tasks involve highly dynamic scene transitions [49, 50]. Consequently, generating high-fidelity future in navigation scenarios via few-step denoising is inherently difficult, posing significant challenges for real-world deployment. While video generation-based methods have been explored in autonomous driving [10, 52], they still suffer from high inference latency, requiring tens of seconds to minutes for generation.

To address this efficiency bottleneck, we adapt PCM [39] to flow-matching paradigm for distilling history-injected I2V model. In this setup, we designate the history-injected I2V model as the teacher model and initialize an architecturally identical student model with the same weights. We then partition the noise schedule into 4 phases, where the student model learns to predict the solution point of each phase on the probability flow ODE [32, 33] trajectory of the teacher model [39]. By minimizing the consistency loss between adjacent timesteps, we progressively distill the inference steps from $N = 50$ to $M = 4$.

F. Stage 4: Action Learning

Drawing inspiration from VPP [14], we freeze the distilled I2V model and employ an inverse dynamics paradigm to predict continuous actions. Specifically, the generated sparse future and the language instruction are injected into the DiT-based action head via cross-attention. Nevertheless, we observe a clear visual discrepancy between the generated and the original ground-truth future frames, which leads to a misalignment between the synthesized dynamics and the original action labels. To combat this inconsistency, we employ DA3 to relabel the generated future frames, thereby ensuring the action supervision is precisely aligned.

We adopt DDIM [31] to reconstruct the relabeled actions $\overline{a}_0$ from noised action $\overline{a}_k = \sqrt{\overline{\beta}_k}\overline{a}_0 + \sqrt{1-\overline{\beta}_k}\epsilon$, where ϵ denotes white noise and $\overline{\beta}_k$ is the noisy coefficient at timestep k. The training process can be formulated as learning a denoised D_ψ:

$$\mathcal{L}_{action}(\psi; A) = \mathbb{E}_{\overline{a}_0,\epsilon,k}||D_\psi(\overline{a}_k, l, \overline{V}) - \overline{a}_0||^2, \quad (5)$$

where $\overline{V}$ denotes the generated sparse latent future.

IV. Evaluations

To validate the efficacy of SparseVideoNav, we design our experiments to investigate: **1)** Can SparseVideoNav, as

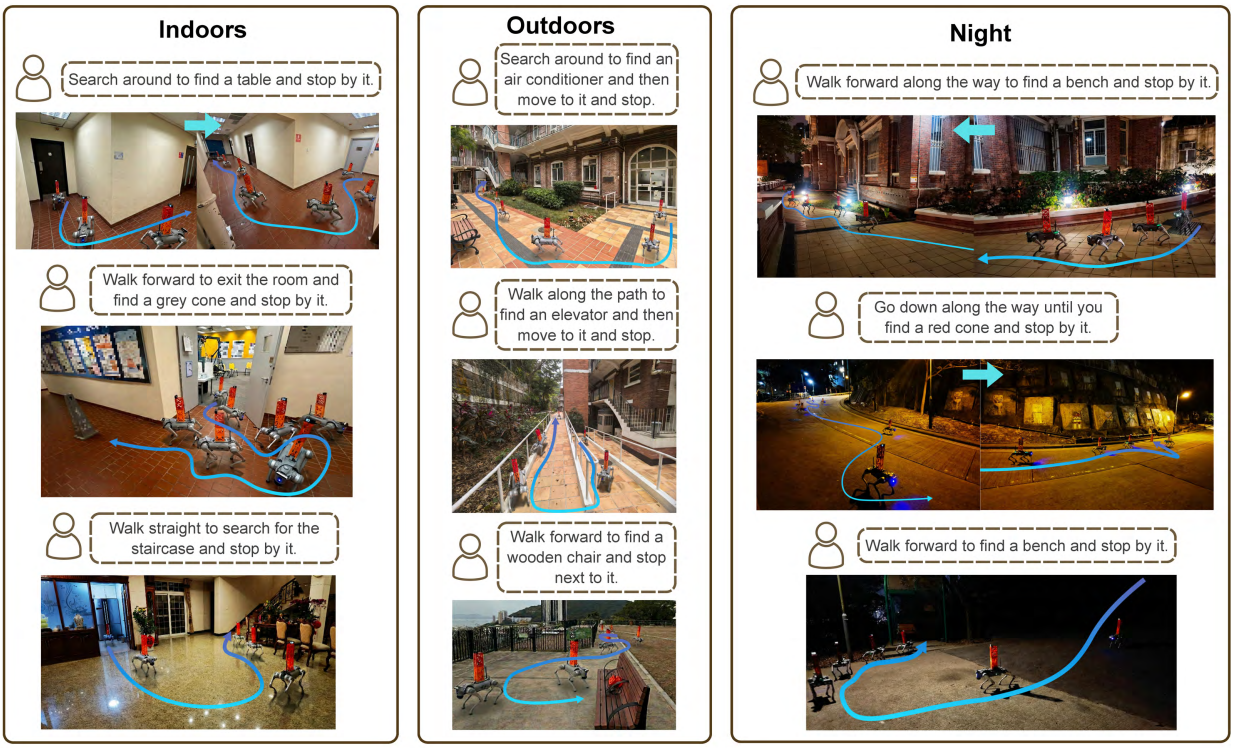

Fig. 5: **Qualitative results of zero-shot beyond-the-view navigation in challenging, unstructured environments.** SparseVideoNav successfully navigates through challenging scenarios, including dead ends, narrow accessible ramp, and hillside with high inclination angles.

a framework grounded in video generation, exhibit foundational competency in following language instructions? (See Sec. IV-B); **2)** Does the sparse foresight successfully translate to superior performance in challenging beyond-the-view scenarios? (See Sec. IV-B and Sec. IV-C); **3)** Does our architectural design strikes a good balance between efficiency and performance? (See Sec. IV-C); **4)** Can SparseVideoNav demonstrate clear scalability with increased data, adaptability to dynamic obstacles, and robustness to camera height variations? (See Sec. IV-C and Sec. F)

A. Experimental Setup

Baselines. To demonstrate the superiority of the VGM backbone, we choose three SOTA LLM-based models for comparison: **1)**Uni-Navid [49] uses video-based LLM to unify diverse VLN tasks; **2)**StreamVLN [42] is designed to accelerate the inference through KV cache. **3)**InternVLA-N1 [35] uses the video generation objective to co-train the backbone.

Evaluation Protocols. To assess the navigation capabilities in real-world environments, we select six diverse unseen scenes across three categories: **Indoors** (Room, Lab Building), **Outdoors** (Yard, Park), and **Night** (Square, Mountain) to test the **zero-shot** generalization capability. For each scene, we design four distinct navigation tasks, consisting of two standard instruction-following navigation (IFN) tasks and two challenging beyond-the-view navigation (BVN) tasks. Detailed specifications for all the tasks are deferred in Sec. B. To ensure statistical reliability, each model is tested 10 times per task, resulting in a total of **240 trials** per model. We report the **Success Rate** averaged across task types for each scene. We observe that LLM-based baselines often stop with a lateral orientation towards the target, whereas our model typically stops facing the target directly. To ensure a fair comparison, we define success solely based on proximity, where a trial is considered successful if the agent stops within 1.5 meters of the target. To minimize the influence of time and weather, we conduct comparative tests for all models within the same time window to ensure consistent environmental conditions. Hardware deployment details are deferred in Sec. E.

To assess foundational language understanding capabilities in the established benchmark, we leverage the VLN-CE R2R [18]. However, we observe that the poor rendering quality of Habitat [28] introduces a severe visual gap, as shown in Fig. 4. Unlike LLM-based methods that map observations into abstract **semantic** spaces, VGMs are intrinsically more sensitive to visual fidelity as they rely on **pixel-level reconstruction**. Since Wan is pre-trained on high-fidelity videos [38], forcing it to reconstruct low-quality scenes would constrain its pre-trained priors. While it is difficult to completely eliminate the influence of these simulation artifacts, we aim to mitigate this asymmetric disadvantage for a fairer evaluation. We carefully inspect all scenes in val-unseen split and isolate a high-fidelity subset (*e.g.*, several hundreds of testing trajectories in total) that holds relatively higher visual quality. We re-run all baseline on this subset, providing more balanced results on their instruction-following performance.

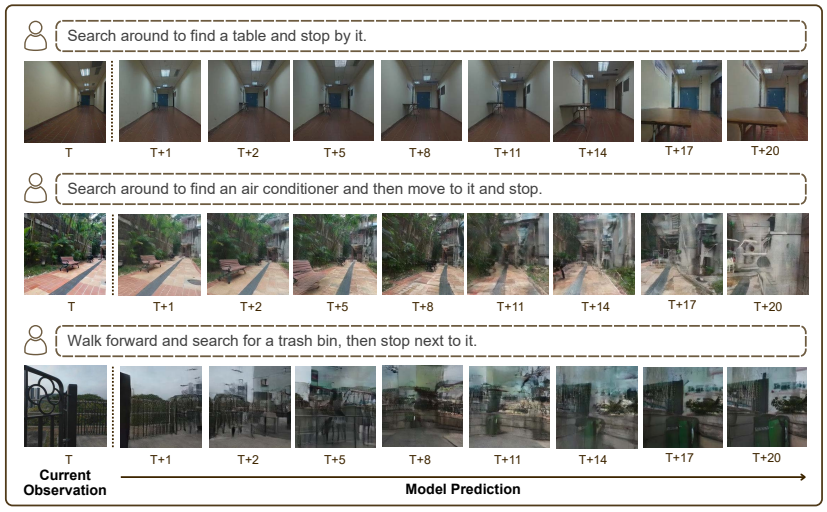

Fig. 6: **Analysis of video generation results** during zero-shot deployment in beyond-the-view navigation.

Fig. 7: **Ablation study on diffusion distillation (Stage 3).**

B. Main Results

As shown in Table I, SparseVideoNav achieves unanimously SOTA zero-shot performance across all real-world scenes on both IFN and BVN tasks. In high-fidelity real-world scenes, the true potential of VGM can be fully unleashed. Compared to the strongest baseline StreamVLN, SparseVideoNav achieves a significant performance improvement, with the average success rate by **+15.0%** on IFN and **+15.0%** on challenging BVN. Notably, the diminished visibility in challenging night environments exacerbates the short-sighted limitation, precipitating a systematic failure across all established baselines on BVN tasks. In contrast, by leveraging robust long-horizon guidance, SparseVideoNav stands as the sole method capable of navigating to these distant goals in such extreme environments.

As shown in Table II, SparseVideoNav achieves competitive performance on the high-fidelity subset of VLN-CE R2R. SparseVideoNav exhibits the smallest discrepancy between success rate and success weighted by path length among all baselines. This minimal gap underscores the profound advantage brought by the long-horizon foresight when guiding the action generation.

Qualitative Results. To further showcase the superiority of SparseVideoNav in BVN, we visualize several trajectories in Fig. 5. SparseVideoNav successfully navigates through these challenging scenarios, including dead ends, narrow accessible ramp, and hillside with high inclination angles.

Analysis. To better understand why SparseVideoNav excels in BVN tasks, we present model prediction results (See Fig. 6) during deployment and several representative failure cases of prior methods (See Fig. 1). As LLM-based baselines rely on short-horizon supervision, they suffer from inherent short-sighted limitations, leading to unexpected turning under long-range uncertainty and premature trapping in dead ends. Conversely, by integrating this guidance with closed-loop feedback, SparseVideoNav can effectively mitigate this limitation. To further inspire future explorations, we detail failure cases of SparseVideoNav in Sec. A.

C. Ablation Study

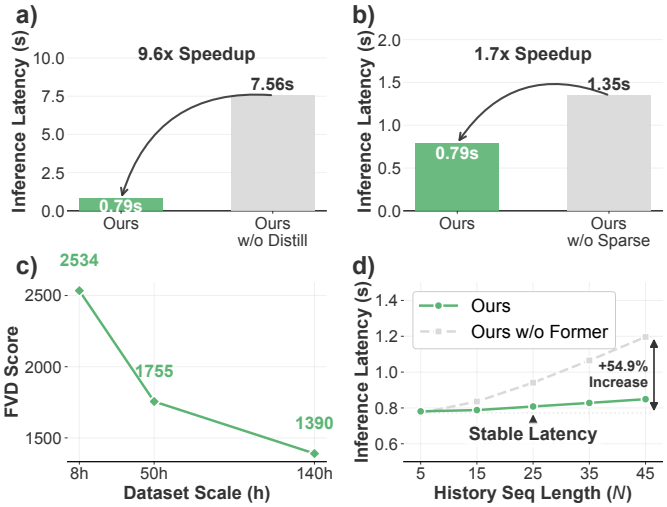

Fig. 8: **Ablation study** on a) data scalability and computational overhead comparison over b) sparse design, c) distillation, and d) history compression.

Diffusion Distillation. As evidenced by Fig. 7, distillation enables VGM to match the visual fidelity of the original 50-step inference with only 4 denoising steps. Crucially, as illustrated in Fig. 8(a), the inclusion of distillation yields a substantial computational advantage, accelerating inference by approximately 10× with slight performance compromise (See *Ablation Study* of Table I variant (c)).

Sparse Video Generation. To demonstrate the advantage of our proposed sparse design, we devise the following variants for comparative analysis:

a) *w distilled 4 steps cont. 2* denotes distilled 4-step variant for generating 2 continuous chunks. Variant (a) is specifically

designed to emulate the effects of short-sighted limitations typically observed in LLM baselines.

b) *w distilled 4 steps cont. 10* denotes distilled 4-step variant for generating 10 continuous chunks.

c) *w/o distilled 50 steps cont. 20* denotes undistilled 50-step variant for generating 20 continuous chunks, where variant (c) can be seen as an oracle compared to SparseVideoNav.

d) *SparseVideoNav w/o Former* denotes our method without Q-Former and Video-Former.

As shown in the *Ablation Study* of Table I, constrained by the same short-horizon supervision issue as LLM baselines, variant (a) yields suboptimal performance. While extending the horizon in variant (b) offers partial mitigation, a distinct performance discrepancy remains compared to SparseVideoNav.

In further comparison with variant (c), SparseVideoNav demonstrates substantial efficiency gains, specifically a $1.7\times$ increase in inference speed (See Fig. 8(b)) and a $1.4\times$ reduction in the cumulative convergence time across training stages 1 and 2 (See Fig. 1). These results show a favorable efficiency-performance trade-off with minimal compromise.

Data Scalability. We conduct the whole training process except Stage 4 on varying data scales: 8h, 50h, and 140h, with additional 3h of unseen data as a validation set to compute FVD [37]. As shown in Fig. 8(c), the consistent downward trend of the FVD curve highlights the scalability of SparseVideoNav, demonstrating the capability to absorb large-scale real-world navigation data.

History Compression Strategy. Fig. 8(d) presents the inference latency across varying history lengths. Former structure decouples latency from history length, ensuring stable inference latency. In contrast to the variant without Former, which incurs a $+54.9\%$ latency penalty at $N = 45$, SparseVideoNav achieves superior efficiency and slightly better performance (See *Ablation Study* of Table I variant (d)).

Pretraining from T2V to I2V. To validate the necessity of Stage 1, we benchmark the training time required for convergence under two variants: **a)** our proposed progressive adaptation; **b)** training Stage 2 directly from scratch. Experiments are conducted on a cluster of 32 NVIDIA H200 GPUs. The results show that variant (b) requires 64 hours to reach convergence, whereas our progressive adaptation strategy converges in 32 hours. This yields a $2\times$ training speedup with similar performance, demonstrating that Stage 1 serves as an efficient initialization to significantly reduce the overall computational overhead.

V. Conclusion and Limitation

We introduce SparseVideoNav, pioneering the use of VGM as the primary backbone for action generation in VLN. By shifting to long-horizon sparse foresight, it overcomes short-sightedness and reduces computational overhead, significantly outperforming SOTA LLM baselines in zero-shot real-world evaluations. Despite its promise, limitations exist: **1)** Our 140-hour dataset is not web-scale [1, 26], requiring future scaling. **2)** Inference latency slightly trails LLM paradigms [42], necessitating accelerated distillation and quantization for VGM.

References

[1] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation World Models. In *CVPR*, 2025. 3, 4, 8, 12

[2] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A Unified Latent Action World Model. In *CVPR*, 2026. 3

[3] Qingwen Bu, Jia Zeng, Li Chen, Yanchao Yang, Guyue Zhou, Junchi Yan, Ping Luo, Heming Cui, Yi Ma, and Hongyang Li. Closed-Loop Visuomotor Control with Generative Expectation for Robotic Manipulation. In *NeurIPS*, 2024. 2

[4] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems. In *IROS*, 2025. 5

[5] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. UniVLA: Learning to Act Anywhere with Task-centric Latent Actions. In *RSS*, 2025. 4

[6] Yihan Cao, Jiazhao Zhang, Zhinan Yu, Shuzhen Liu, Zheng Qin, Qin Zou, Bo Du, and Kai Xu. CogNav: Cognitive Process Modeling for Object Goal Navigation with LLMs. In *ICCV*, 2025. 2

[7] Li Chen, Chonghao Sima, Kashyap Chitta, Antonio Loquercio, Ping Luo, Yi Ma, and Hongyang Li. Intelligent Robot Manipulation Requires Self-Directed Learning. *Authorea Preprints*, 2025. 2

[8] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged Robot Vision-Language-Action Model for Navigation. In *RSS*, 2025. 1

[9] Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. UniMax: Fairer and More Effective Language Sampling for Large-Scale Multilingual Pretraining. In *ICLR*, 2023. 4

[10] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. MagicDriveDiT: High-resolution Long Video Generation for Autonomous Driving with Adaptive Control. *arXiv preprint arXiv:2411.13807*, 2024. 3, 5

[11] Zeying Gong, Rong Li, Tianshuai Hu, Ronghe Qiu, Lingdong Kong, Lingfeng Zhang, Yiyi Ding, Leying Zhang, and Junwei Liang. Stairway to Success: Zero-Shot Floor-Aware Object-Goal Navigation via LLM-Driven Coarse-to-Fine Exploration. *arXiv preprint arXiv:2505.23019*, 2025. 2

[12] Noriaki Hirose, Amir Sadeghian, Marynel Vázquez, Patrick Goebel, and Silvio Savarese. GoNet: A Semi-

Supervised Deep Learning Approach for Traversability Estimation. In *IROS*, 2018. 4

[13] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable Autonomous Control for Social Navigation. 2023. 4

[14] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations. In *ICML*, 2025. 2, 3, 4, 5, 13

[15] Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards Unified Latent VLA for Whole-Body Loco-Manipulation Control. In *ICLR*, 2026. 4

[16] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An Open-Source Vision-Language-Action Model. In *CoRL*, 2024. 4, 5

[17] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014. 4

[18] Jacob Krantz and Stefan Lee. Sim-2-Sim Transfer for Vision-and-Language Navigation in Continuous Environments. In *ECCV*, 2022. 6

[19] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In *EMNLP*, 2020. 1

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023. 3, 4, 13

[21] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal World Modeling for Robot Control. *arXiv preprint arXiv:2601.21998*, 2026. 3

[22] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the Visual Space from Any Views. *arXiv preprint arXiv:2511.10647*, 2025. 4, 12, 15

[23] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *ICLR*, 2023. 4

[24] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. InstructNav: Zero-shot System for Generic Instruction Navigation in Unexplored Environment. In *CoRL*, 2024. 2

[25] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 13

[26] Yanghong Mei, Yirong Yang, Longteng Guo, Qunbo Wang, Ming-Ming Yu, Xingjian He, Wenjun Wu, and Jing Liu. UrbanNav: Learning Language-Guided Urban Navigation from Web-Scale Human Trajectories. In *AAAI*, 2026. 8

[27] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *ICCV*, 2023. 13

[28] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019. 3, 6

[29] Dhruv Shah, Benjamin Eysenbach, Nicholas Rhinehart, and Sergey Levine. Rapid Exploration for Open-World Navigation with Latent Goal Models. In *CoRL*, 2021. 4

[30] Modi Shi, Li Chen, Jin Chen, Yuxiang Lu, Chiming Liu, Guanghui Ren, Ping Luo, Di Huang, Maoqing Yao, and Hongyang Li. Is Diversity All You Need for Scalable Robotic Manipulation? *arXiv preprint arXiv:2507.06219*, 2025. 5

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021. 5

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*, 2021. 5

[33] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. In *ICML*, 2023. 5

[34] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. VEEGAN: Reducing Mode Collapse in Gans Using Implicit Variational Learning. In *NeurIPS*, 2017. 4

[35] InternVLA-N1 Team. InternVLA-N1: An Open Dual-System Navigation Foundation Model with Learned Latent Plans, 2025. URL https://internrobotics.github.io/internvla-n1.github.io/static/pdfs/InternVLA_N1.pdf. 2, 3, 5, 6, 12, 14

[36] Hoang Thanh-Tung and Truyen Tran. Catastrophic Forgetting and Mode Collapse in GANs. In *IJCNN*, 2020. 4

[37] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A New Metric for Video Generation, 2019. URL https://openreview.net/forum?id=rylgEULtdN. 8

[38] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 4, 6, 13, 15

[39] Fu-Yun Wang, Zhaoyang Huang, Alexander Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased Consistency Models. In *NeurIPS*, 2024. 5

[40] Shaoan Wang, Jiazhao Zhang, Minghan Li, Jiahang Liu, Anqi Li, Kui Wu, Fangwei Zhong, Junzhi Yu, Zhizheng Zhang, and He Wang. TrackVLA: Embodied Visual Tracking in the Wild. *arXiv preprint arXiv:2505.23189*, 2025. 2

[41] Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, et al. Ground Slow, Move

Fast: A Dual-System Foundation Model for Generalizable Vision-and-Language Navigation. In *ICLR*, 2026. 1

[42] Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. StreamVLN: Streaming Vision-and-Language Navigation via Slowfast Context Modeling. In *ICRA*, 2026. 1, 2, 5, 6, 8

[43] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 Technical Report. *arXiv preprint arXiv:2511.18870*, 2025. 2, 4

[44] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. In *ICLR*, 2024. 5

[45] Jiazhi Yang, Kashyap Chitta, Shenyuan Gao, Long Chen, Yuqian Shao, Xiaosong Jia, Hongyang Li, Andreas Geiger, Xiangyu Yue, and Li Chen. ReSim: Reliable World Simulation for Autonomous Driving. In *NeurIPS*, 2025. 3

[46] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World Action Models are Zero-shot Policies. *arxiv preprint arXiv:2602.15922*, 2026. 2

[47] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. In *ICRA*, 2024. 2

[48] Bangguo Yu, Yuzhen Liu, Lei Han, Hamidreza Kasaei, Tingguang Li, and Ming Cao. VLN-Game: Vision-Language Equilibrium Search for Zero-Shot Semantic Navigation. *arXiv preprint arXiv:2411.11609*, 2024. 2

[49] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks. In *RSS*, 2025. 1, 2, 3, 5, 6

[50] Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied Navigation Foundation Model. In *ICLR*, 2026. 1, 3, 5

[51] L Zhang, X Hao, Q Xu, Q Zhang, X Zhang, P Wang, J Zhang, Z Wang, S Zhang, and R Xu. A Novel Memory Representation via Annotated Semantic Maps for VLM-Based Vision-and-Language Navigation. In *ACL*, 2025. 2

[52] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation. In *AAAI*, 2025. 3, 5

APPENDIX

In the Appendix, we first compile "motivating" questions in Sec. A. Task specifications details in Sec. B are presented to supplement Sec. IV-A in the main paper. We then provide additional elaborations on data curation details, implementation details and hardware deployment details in Sec. C, Sec. D and Sec. E respectively. Next, we discuss the adaptability and robustness of SparseVideoNav in Sec. F. The Appendix concludes with Sec. G, which details the ethics statement and licenses.

A. Motivating Questions

Q1: *What are the detailed failure cases of SparseVideoNav?*

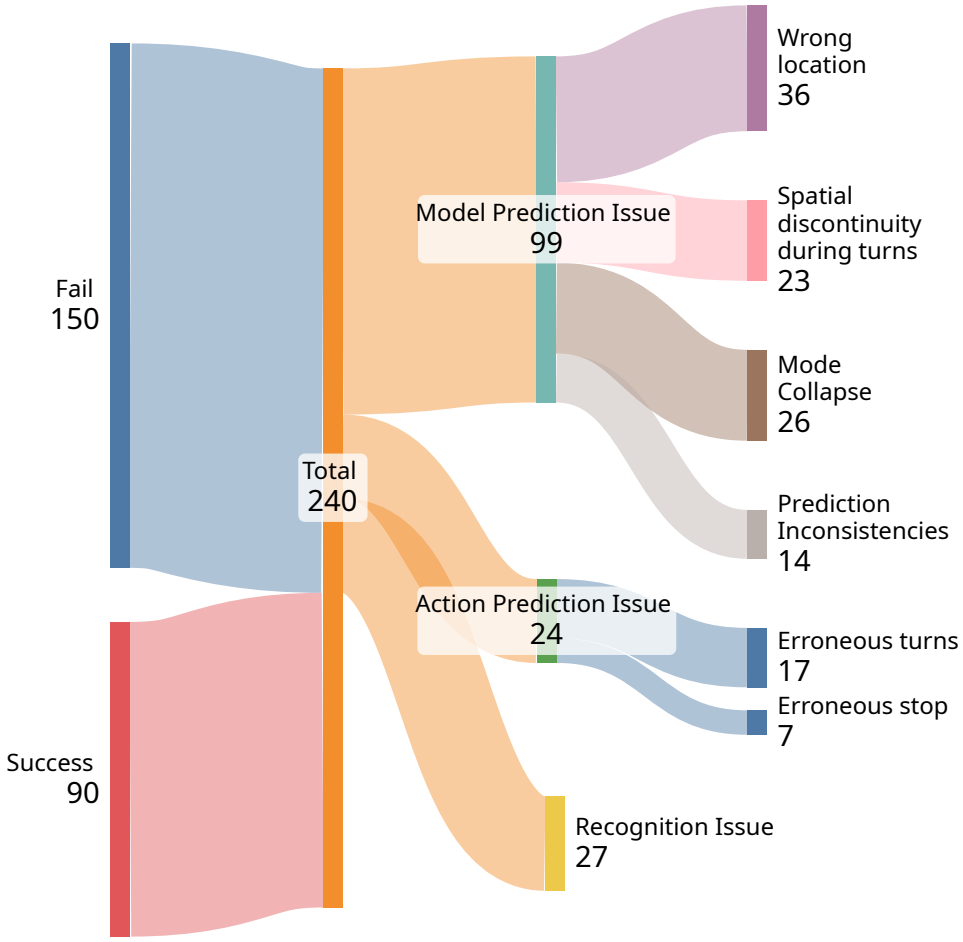

Fig. A-1: **Detailed analysis of the failure cases of SparseVideoNav.**

We conduct a comprehensive statistical analysis of the failure cases observed across all 240 real-world evaluation trajectories. As illustrated in Fig. A-1, out of 150 failure instances, we categorize them into three primary groups:

- **Model Prediction Issues (99 cases):** This category encompasses failures stemming from the video generation process. The most prevalent error is wrong location (36 cases). This typically occurs in expansive, open environments where the target position "imagined" by the generative model misaligns with the real world (*e.g.,* early turning at a certain crossing). The agent ultimately fails to localize the true destination within the context limits. The second factor is mode collapse (26 cases, occurring more frequently in night scenes), which we attribute to the currently limited training data volume (approximately 140 hours). Additionally, spatial discontinuity during turns (23 cases) occurs due to the constraints of the fixed sparse generation interval. Finally, prediction inconsistencies (14 cases) arise when the navigation intent predicted at a certain timestep misaligns with subsequent predictions during the BVN task. These observations strongly motivate our future efforts in exploraing lifelong navigation, scaling the dataset volume and introducing adaptive sparsity mechanisms to better anchor predictions in complex open worlds.
- **Action Prediction Issues (24 cases):** This group involves misalignments between the predicted visual future and the generated actions. Specific manifestations include erroneous turns (17 cases)—such as visually predicting a right turn but

outputting a left-turn command—and erroneous stops (7 cases), where a visually predicted slight turn prematurely triggers a stop action. We hypothesize that this stems from our decoupled design choice (further discussed in Q2). While decoupling ensures critical training stability for the video generation model, relying on a separately trained lightweight action head may limit its generalization capabilities. Future work exploring an effective vision-action joint training architecture in embodied navigation could potentially alleviate this gap.

- **Recognition Issues (27 cases):** In some instances, VGM fails to identify the target object category, even when situated in close proximity. We attribute this representational shortcoming to the relatively small parameter size (1.3B) of our chosen base model. Future directions to resolve this include leveraging larger-scale base models coupled with advanced quantization and distillation algorithms to maintain inference efficiency.

Q2: *What is the structural necessity of the four-stage training pipeline?*

Decoupling the video model training (Stage 2) and action training (Stage 4) is **a pragmatic design choice prioritizing training stability**. We initially considered vision-action joint optimization. However, unlike robotic manipulation where the camera view remains largely static and even 1-step denoising can extract sufficient features, real-world navigation involves drastic scene variations. In such dynamic settings, even performing 10+ denoising steps struggles to extract usable representations, making simultaneous optimization for both high-fidelity video generation and precise action prediction highly susceptible to instability. To ensure robust learning of the foundational video representations, we opt for this decoupled paradigm as an effective first step.

As for adaptation (Stage 1) and distillation (Stage 3), they are served for computational efficiency (See Sec. IV-C), which do not introduce great complexity.

Q3: *What is the distinction between Navigation World Model [1] and SparseVideoNav?*

To the best of our knowledge, SparseVideoNav is the first to apply a video generation model as the primary backbone for action generation in VLN. The distinction between SparseVideoNav and prior Navigation World Models stems from a fundamental modality mismatch:

- **Navigation World Model**: It is strictly **action-conditioned**. The architecture relies on low-level action inputs to predict the immediate next frame, rather than generating long-horizon videos.
- **SparseVideoNav**: Conversely, our framework is **language-conditioned**. It directly processes high-level language intents to achieve long-horizon video foresight, completely eliminating the need for action inputs.

Furthermore, we demonstrate that utilizing longer sequence supervision during training yields notable performance improvements (See Sec. IV-C). This effectively validates the significance of sparse video generation in effectively extending the prediction span.

Q4: *What about the object and language diversity of our collected real-world dataset?*

Regarding object diversity, we aim for open-world navigation. The targets extend far beyond "benches or chairs" in common simulation benchmarks, encompassing diverse **authentic physical entities** (e.g., fire hydrants, vehicles, boats, fence, elevators, road signs, bus station). While their locations may appear "predictable", they strictly follow natural real-world semantic priors.

Regarding language diversity, we explicitly **depart from** the previous paradigm of dense instructions. Our motivation is to minimize the human interaction burden by using simple and high-level intents.

Q5: *Compared to the pixel-goal guidance in InternVLA-N1 [35], what is the advantage of sparse video guidance?*

The point-goal generated by language models are fundamentally confined to the current view. Consequently, it often fails to assist the agent in escaping when trapped in dead ends. In contrast, video generation models can leverage long-horizon video guidance to imagine a backtracking trajectory to exit dead ends, thereby mitigating this perspective limitation and leading to more robust solutions for BVN tasks.

B. Task Specifications and Details

Fig. A-2 and Fig. A-3 provide the complete specifications for all 24 zero-shot evaluation tasks across six real-world scenes. Each scene contains 2 instruction-following navigation (IFN) tasks and 2 beyond-the-view navigation (BVN) tasks.

C. Data Curation Details

As illustrated in Figure A-4, our pipeline efficiently transforms raw video streams into high-quality training image-action pairs. The process begins with temporal sampling, where raw videos are decimated to 4 FPS to reduce redundancy. We then utilize Depth Anything 3 (DA3) [22] to estimate the 6-DoF camera extrinsics for each frame, reconstructing the 3D trajectory of the agent. Finally, action extraction is performed by calculating the relative pose transformations between frames. These 3D motions are projected onto the local XY plane to derive continuous action labels $(\Delta x, \Delta y, \Delta\theta)$, while static segments and views with extreme pitch angles are filtered out to ensure high-quality selection.

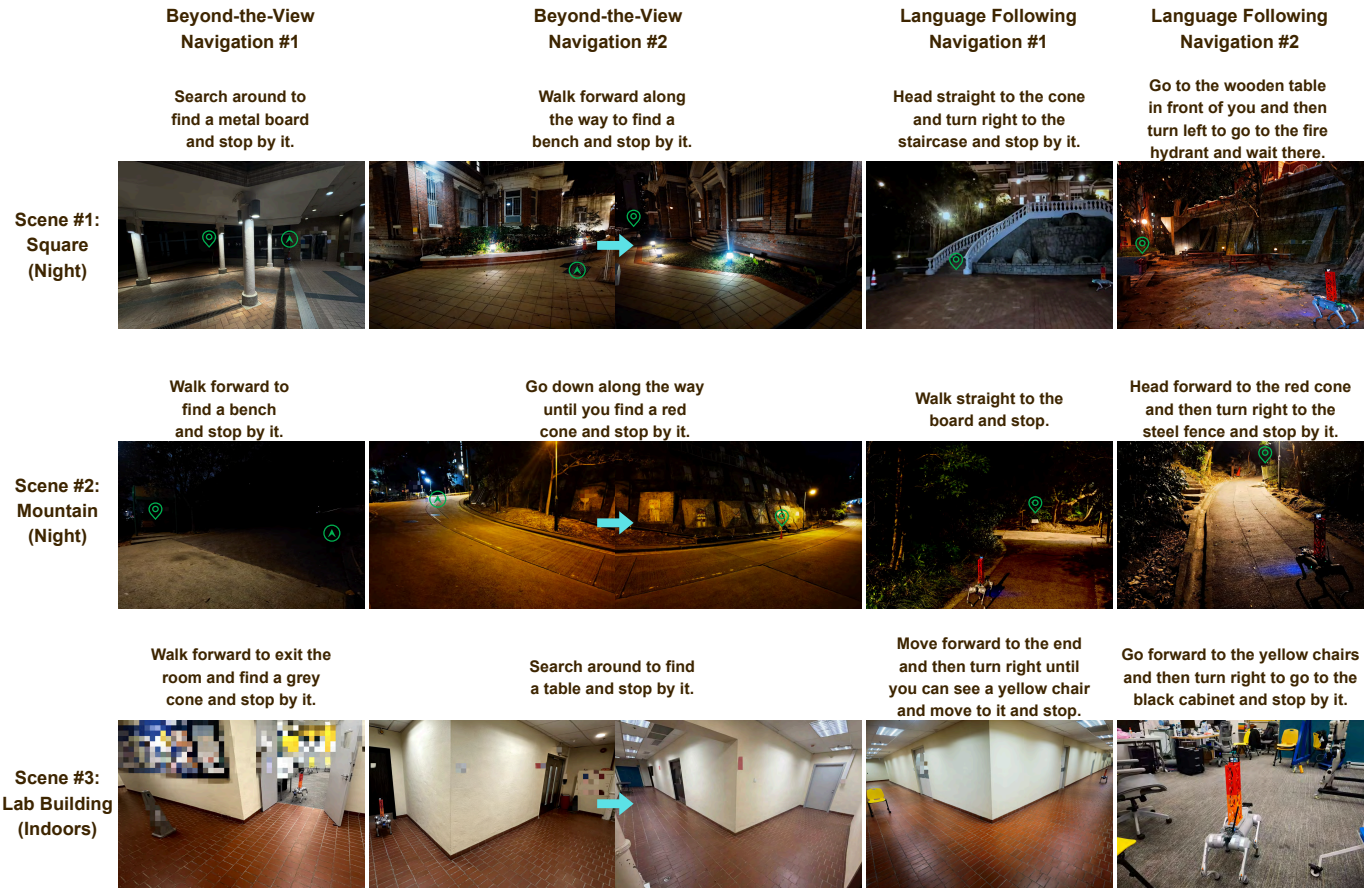

Fig. A-2: **Task specifications.** We present the language instructions, the initial positions of robot dog and the task settings for each scene. When the precise locations can not be clearly seen, we utilize a green arrow icon to denote the initial positions and a location pin to mark the destination for visualization clarity.

TABLE A-I: **Hyperparameter configurations for the four-stage training pipeline.**

Stage	Batch Size	Learning Rate	LR Scheduler	Trainable Params	Sampling Strategy	Discrete Steps
Stage 1	32	1×10^{-5}	Constant	1.3B	FM (Uniform)	1000
Stage 2	32	1×10^{-5}	Constant	1.8B	FM (Uniform)	1000
Stage 3	32	5×10^{-6}	Constant	1.7B	PCM (Distill)	50
Stage 4	256	5×10^{-5}	Cosine	23.4M	DDIM	100

D. Implementation Details

Training Configurations. Our model is trained using the AdamW optimizer [25] on a cluster of 32 NVIDIA H200 GPUs, with the full four-stage pipeline requiring approximately 64 hours to complete. All training hyper-parameters of each stage is presented in Table A-I.

Network Architecture Configurations. Our system integrates three core components: a video generation backbone, a history compression module, and an action prediction head. For the video backbone (Stage 1), we directly adopt the Wan2.1 T2V-1.3B [38] architecture. To handle high-dimensional historical context, we implement a two-stage compression strategy: a Q-Former [20] first reduces temporal redundancy, followed by a Video-Former [14] that performs $4\times$ spatial downsampling. This reduces the token burden to 2,560 while preserving essential context. We present the detailed hyperparameters of Q-Former and Video-Former in Table A-II.

For navigation control, we design an inverse dynamics-based action head that couples a feature-aggregating Video-Former with a Diffusion Transformer (DiT) [27]. Video-Former encodes spatiotemporal features from 640 sparse future latent tokens.

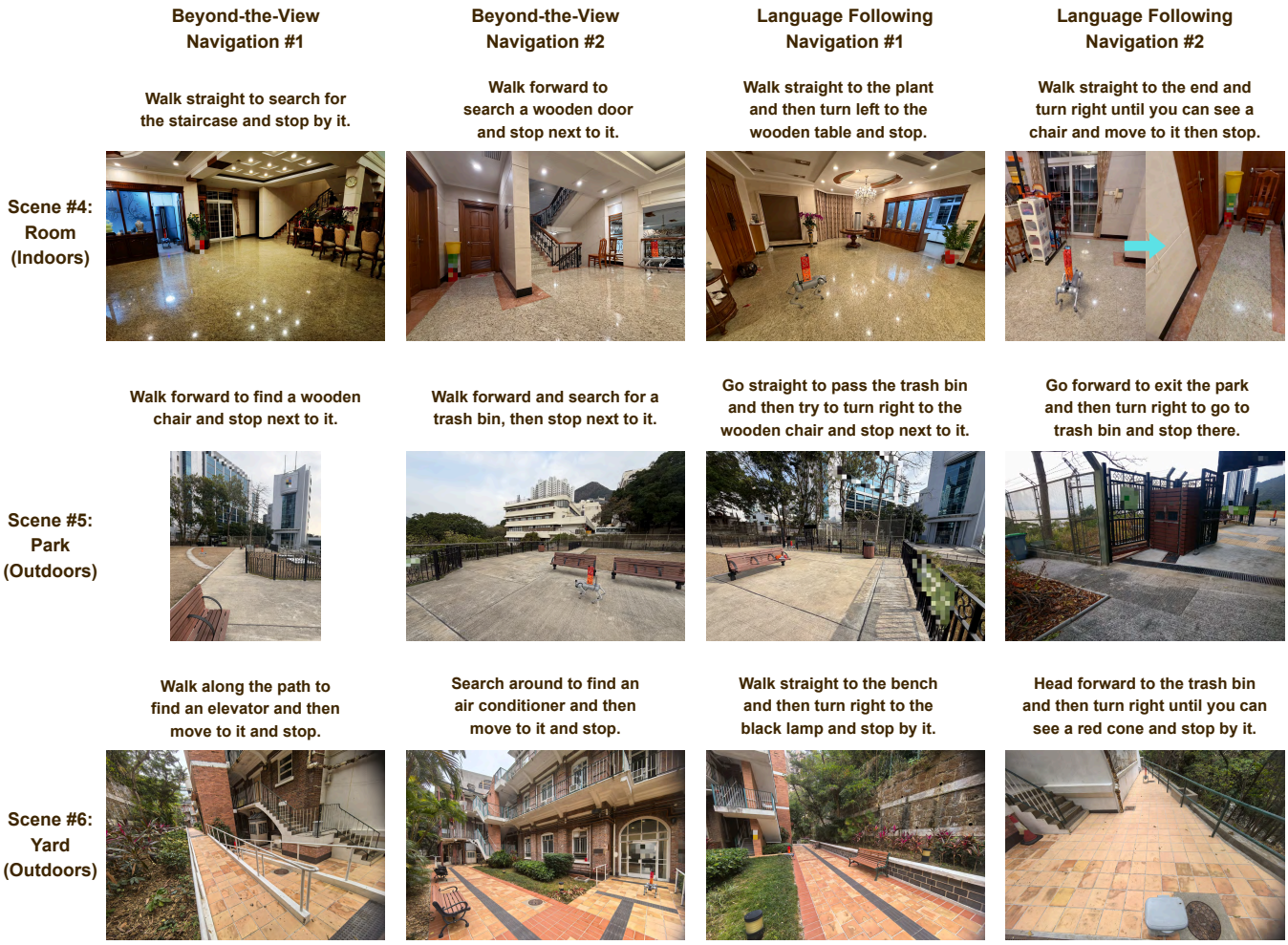

Fig. A-3: **Task specifications.** We present the language instructions, the initial positions of robot dog and the task settings for each scene.

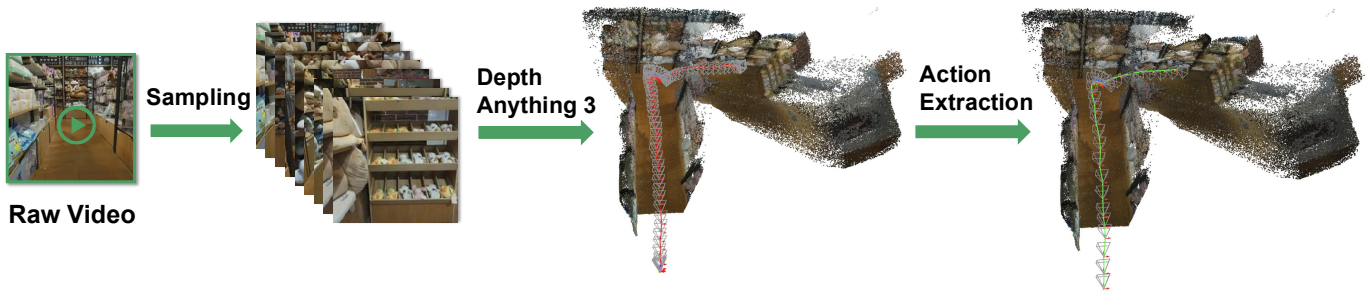

Fig. A-4: **Data Curation Pipeline.** The workflow consists of three stages: (1) temporal sampling from raw video; (2) pose estimation via Depth Anything 3; and (3) extrinsic-based action extraction.

These features then serve as the cross-attention context for the DiT to predict continuous 8-step action trajectories. Detailed parameters for these modules are summarized in Table A-III.

E. Hardware Deployment Details

We perform real-world experiments based on Unitree Go2 robotic dog. The robot is equipped with an upward facing camera DJI Osmo Action 4 for stabilized RGB observations. However, given that InternVLA-N1 [35] requires depth input, we adhere to its original configuration by employing an Intel® RealSense™ D455 camera, mounted with a 15° downward pitch to

TABLE A-II: **Hyper-parameters of Q-Former and Video-Former used for history compression.**

Q-Former Configuration		Video-Former Configuration	
Hidden Dim	512	Hidden Dim	512
Layers	4	Layers	6
Heads	8	Heads	8
Num. Latents	10240	Num. Latents	2560

TABLE A-III: **Hyper-parameters of the inverse dynamics-based action head**, including the Action Video-Former and the Diffusion Transformer.

Action Video-Former		Diffusion Transformer	
Hidden Dim	256	Hidden Dim	256
Layers	8	Layers	12
Heads	8	Heads	8
Num. Latents	640	Num. Actions	8

acquire RGB-D observations. The camera is secured to the back of Go2 using a custom 3D-printed bracket, positioned at a height of approximately 1m above the ground. This mounting height is standardized across all evaluated methods. We deploy SparseVideoNav and all the baselines on a remote workstation with an RTX 4090 GPU. The Go2 robot continuously sends visual observations to the server. Upon receiving the images, the server processes them and generates corresponding navigation action commands, which are then sent back to Go2 for execution.

F. Additional Experiments

Dynamic Pedestrians Avoidance. Given that DA3 [22] struggles with reliable action estimation in the presence of frontal dynamic pedestrians, we filter out such trajectories to preserve the accuracy of the actions. Remarkably, despite this exclusion, SparseVideoNav exhibits an emergent capability to dynamically avoid pedestrians during deployment, underscoring its strong adaptability and generalization, as shown in Fig. A-5.

Camera Height Insensitivity. We further observe that while LLM-based paradigm is notably susceptible to camera height shifts, video generation-based paradigm maintains robust against these discrepancies. Despite being trained on data collected at an approximate height of 1m, SparseVideoNav successfully performs navigation with a fixed camera height of 50cm, as shown in Fig. A-6.

G. Data Ethics and License

Privacy Protection and De-identification. To strictly protect individual privacy, we implement a robust and irreversible de-identification processing pipeline. All captured video frames are processed by human expert to identify all sensitive information. All sensitive information, including human faces and vehicle identifiers, is permanently blurred before being used for model training or stored in our servers. We ensure that no identifiable personal information is present in the final dataset.

Data Collection Ethics. Our data collection involves human operators recording in diverse public environments. All operators are briefed on ethical guidelines, ensuring that they follow local regulations and avoid recording in restricted or private areas without explicit authorization. Data collection is performed using a handheld DJI Osmo Action 4 in real-world environments. This manual approach ensures data diversity and procedural transparency throughout the recording process.

License. We build upon Wan2.1 [38], which is under Apache License 2.0. The code and dataset for SparseVideoNav would be open-sourced under CC BY-NC-SA 4.0 License.

Fig. A-5: **Discussion of SparseVideoNav under dynamic pedestrian interference.** SparseVideoNav successfully avoids the oncoming pedestrian and reaches the door.

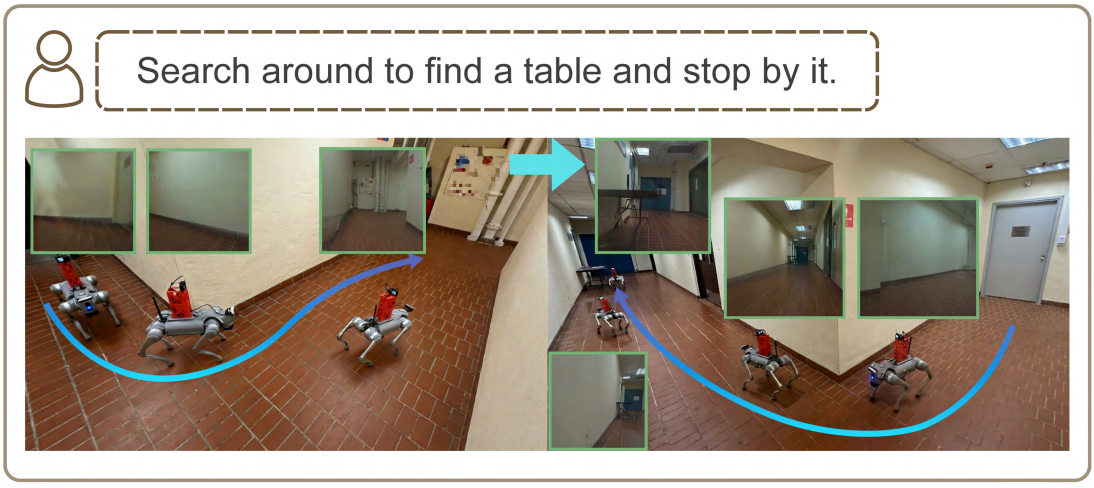

Fig. A-6: **Discussion of SparseVideoNav with the camera fixed at 50cm.** Images in green boxes demonstrate the predictions generated by SparseVideoNav. SparseVideoNav exhibits strong robustness against camera height variations.